Reasoning across multilingual learning resources in human genetics

Conference or Workshop Item

Reasoning across Multilingual Learning Resources in Human Genetics

Zdenek Zdrahal, Petr Knoth, Trevor Collins, Paul Mulholland

Knowledge Media Institute, The Open University, Milton Keynes, UK

Key words *semantic annotation, learning resources, multilingual ontology, genetics, semantic similarity, automatic e-course generation, multilingual educational environments*

Abstract:

This paper describes how semantic annotations in terms of a domain ontology and theme hierarchy can be used for organising and reusing educational resources. A case study is presented in the domain of human genetics. The technology has been developed as a part of the Eurogene project and allows the user to submit, annotate and retrieve multimedia learning resources in nine European languages. We present two use case examples: Query by example and discovering learning pathways.

1 Introduction

Eurogene is a 36 month education oriented project supported by the Commission of European Communities (CEC). The objectives of Eurogene are to develop methods for the sharing of and reasoning across learning resources in human genetics. The project consortium consists of 21 partners, 16 of which are academic content providers and users from 11 European countries, two are specialised in machine translation of natural languages and three are responsible for the project infrastructure, knowledge technologies and quality assurance. The support for the collaboration of learners and content providers is provided in nine European languages: English, French, German, Italian, Spanish, Dutch, Greek, Czech and Lithuanian. The level of support differs due to the differences in language technologies available for individual languages.

This paper describes results achieved within the first two years.

The content partners can play two roles in the knowledge sharing model:

- as content providers they submit their educational material to the Eurogene repository and take part in the process of semantic annotation,
- as content users they query the repository to acquire relevant learning resources organised according to the educational goal. It is expected that some content users, e.g. university lecturers, are also content providers though the repository might be used also by students who are only content users.

Semantic annotations of learning resources makes it possible to support various learning scenarios, including the following:

1. A student reads the learning material and is looking for another resource with a similar theme in the same or a different language. She submits the original learning material as a query and expects the system to provide alternatives. This model is called query-by-example (QBE).

2. A lecturer defines the prerequisites for the course, the expected knowledge to be acquired by the student at the end of the course and the languages the students understand. The system offers a sequence of presentations that guides the student through the "knowledge space" from the start to the end point.
3. For sequentially organised course materials the lecturer/learner can ask for alternate materials that explain the presented themes in more or less detail (and/or in different languages).

We approach this class of problems by annotating resources using domain specific knowledge structures which are used for constructing solutions. Eurogene exploits two different, but interconnected knowledge structures: a multilingual genetic ontology and a hierarchy of genetic themes/topics. The ontology consists of fundamental domain concepts and their relationships. The theme/topic hierarchy represents domain decomposition at a higher level of abstraction.

2 Educational content

The learning resources can be submitted in a number of different formats. The Eurogene repository allows the content providers to upload PowerPoint presentations, MS Word, PDF documents, plain text, web pages, images, audio or video clips. All textual resources are automatically annotated (unless the text is represented as an image). When a new text resource is submitted it is converted into the PDF format and annotated in terms of concepts from the multilingual domain ontology. The content provider has then the opportunity to confirm or reject each concept proposed by the annotation algorithm. The text is machine translated into up to 8 additional languages that are available as PDF files. The SYSTRAN machine translation software used in the project allows content to be translated from English to French, German, Spanish, Italian, Dutch and Greek, and between these languages though not all pairs (e.g. Spanish to Dutch) are supported. Non-textual resources, such as images, audio or video are annotated manually, but the selection of terms is guided by the ontology.

All resources are also associated with one or more themes in the theme hierarchy. This association is currently carried out manually by selecting from the theme tree. It is expected that as soon as the critical mass of educational content is uploaded and evaluated, this process will be automated and the content provider will only confirm the proposed associations. The submission of a textual resource is shown in Figure 1.

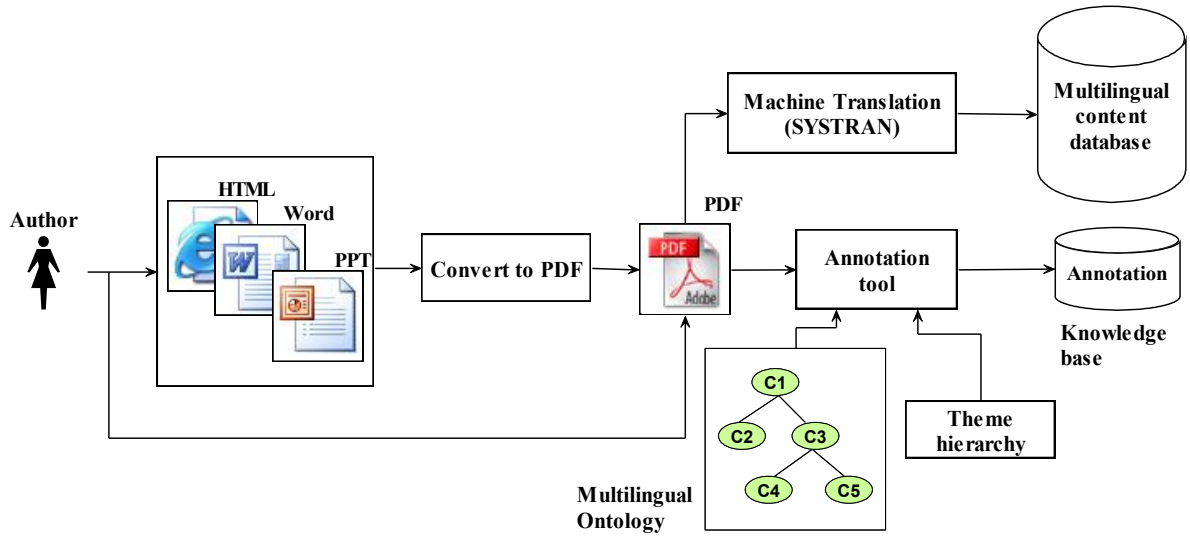

Figure 1. Automatic annotation of textual resources

In addition to the concepts associated automatically with the content, the provider can add his/her own terms as free text. These terms are not associated with the ontology and cannot be

used for reasoning, but can be part of search queries to improve resource retrieval. In addition, they are stored in the database as potential candidates for future ontology extensions.

When a new resource is submitted, the content provider also specifies the target audience. Eurogene distinguishes learners at 6 academic levels: GCSE, A-level, QC1, QC2, QC3 and Expert. GCSE level is intended for students under 16, A-level is for students from 16-18. QC1, QC2 and QC3 correspond to the three cycles in university education as declared by the Bologna process. Roughly speaking, these cycles are bachelor, master and PhD, respectively. The expert level is intended for professionals in the field. The same resource may be classified as suitable for more than one academic level.

3 Ontology and theme/topic hierarchy

Ontology is the core knowledge structure of Eurogene, used for annotating educational content, search and reasoning. When developing the ontology Eurogene partners were constrained by the trade-off between the domain coverage and design economy. There are many highly-specialised, and verified monolingual ontologies related to genetics on the web (e.g. Gene Ontology). However, each of them covers only a small part of the supported domain, and they are not designed for educational purposes. Their integration would be expensive and the translation of relevant parts into 9 different languages would be cost prohibitive. For these reasons, a new Eurogene monolingual ontology was developed which was later translated into 9 European languages. The initial domain conceptualisation was done by selecting six well-established genetic glossaries and merging their content. These glossaries together with the number of genetic concepts they provided are shown in Table 1

Nr	Author/origin	Concepts
1	University of Washington, Seattle	**282**
2	National Institute of General Medical Sciences, Bethesda	**52**
3	Emery's Elements of Medical Genetics	**811**
4	ThinkQuest The Oracle Education Foundation	**182**
5	University of Michigan, The Center for Genetics in Health and Medicine	**76**
6	Centre for Genetics Education in Sydney, Australia	**256**

Table 1. Genetic glossaries

As there was a significant overlap between these glossaries, the initial merged version contained only 1302 concepts. Some glossaries were already structured and represented in accordance with the ISO standard for defining controlled vocabularies [2].

Eurogene distinguishes concepts as semantic objects and terms as their verbal representation. In natural language the same concept can be represented by multiple terms i.e. synonyms. The set of synonyms associated with the same concept is called a synset [7], [11]. For semantic annotation, each concept could be represented by one selected term from its synset. This term is called the preferred term. All other terms from the synset found in the text are represented in the annotation by the preferred term. Using preferred terms to represent whole synset allows search and reasoning algorithms to work with word semantics without imposing any constraint on the original text.

The ISO 2788 norm [2] also allows us to use as preferred terms for annotation words that are not synonyms, e.g. class names for the names of instances (“rocks” as the preferred term for “granite” and “slate”), or quasisynonyms – words that are closely related but with different meaning (e.g. “electric resistance” as the preferred term for “electric conductivity”). However, for ontologies these semantic misrepresentations are unacceptable and have to be corrected. Our multilingual ontology is created by translating concept (and not terms). Consequently, concepts can be described by a different number of terms in different languages. In Eurogene, the SYSTRAN machine translation software first learned the translation of domain dependent

concepts from a selection of genetic documents, then the ontology was translated and finally, the translation was corrected by bi-lingual domain experts. In some cases, the translation of ontology was done directly by the expert.

Additional concepts and relations were acquired by mapping Eurogene concepts to existing ontologies, and filling the gaps. In this process the following ontologies were used Gene Regulation, Mammalian Phenotype, Gene Ontology, NCI Thesaurus and Universal Medical Language System (UMLS) [9]. For example, about 60% of Eurogene concepts were found in UMLS. An example of ontology completion is shown in Figure 2.

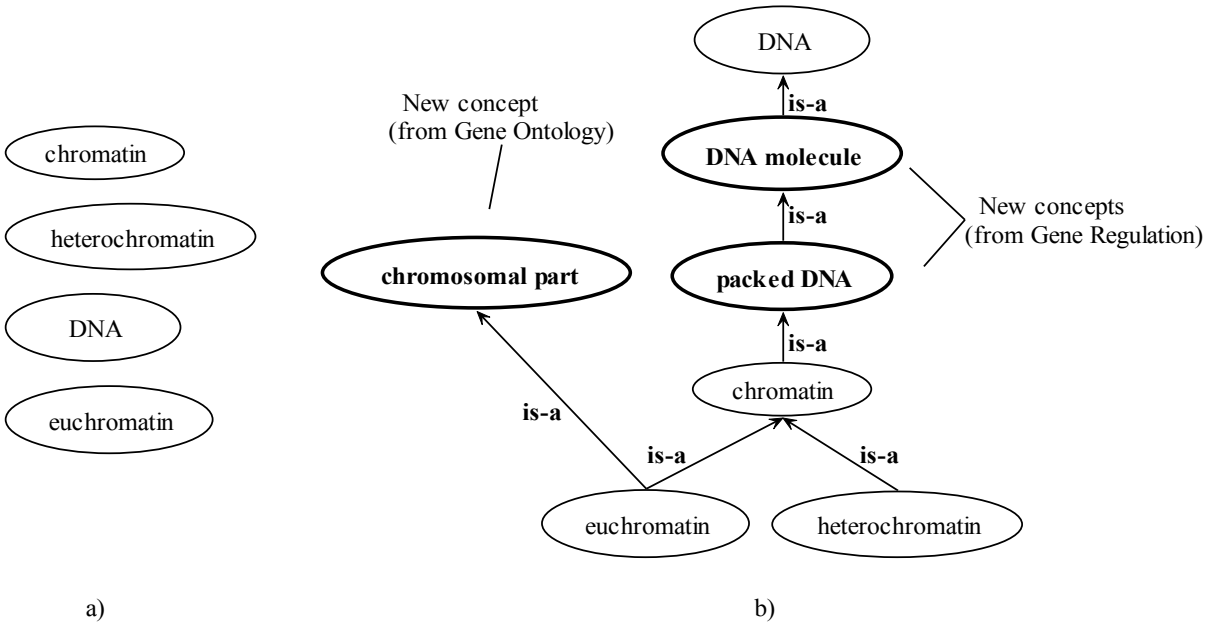

Figure. 2 Ontology completion: three added concepts and their is-a relations

Figure 2 a) shows four concept of the Eurogene ontology before completion. After mapping to the Gene Regulation ontology two new concepts (packed DNA and DNA molecule) and 5 is-a (class – subclass) relations were included. Mapping to the Gene Ontology produced another new concept (chromosomal part) and one new is-a relation.

During the development, the ontology was also updated as new content annotation and reasoning revealed missing concepts. At present, the missing concepts are collected and from time to time evaluated. The concepts are assessed by domain experts and selected ones integrated with the ontology. At present, the multilingual ontology consists of 2,117 concepts described by 9,832 terms in nine languages.

The theme/topic hierarchy shown in Figure 3 was developed in close interaction with experts. In standard knowledge acquisition sessions, groups of domain experts were interviewed, their responses evaluated and the theme hierarchy constructed. In the next session, the results were presented back to the group for comments and corrections. This was repeated until a stable result was achieved. When all trees for all subdomains were completed, they were merged together and duplicity discussed and resolved. The proposed tree was eventually validated by the rest of the community.

The purpose of this hierarchy is to provide a coarse-grain description of the domain that can be used both to speed up search and to prune the search space used for more complex reasoning tasks. The maximum depth of the hierarchy is ten. At present, the content provider associates one or more themes from the hierarchy manually, however experiments have been carried out with calculating correlations between terms concepts selected for content annotation and theme of the presentation.

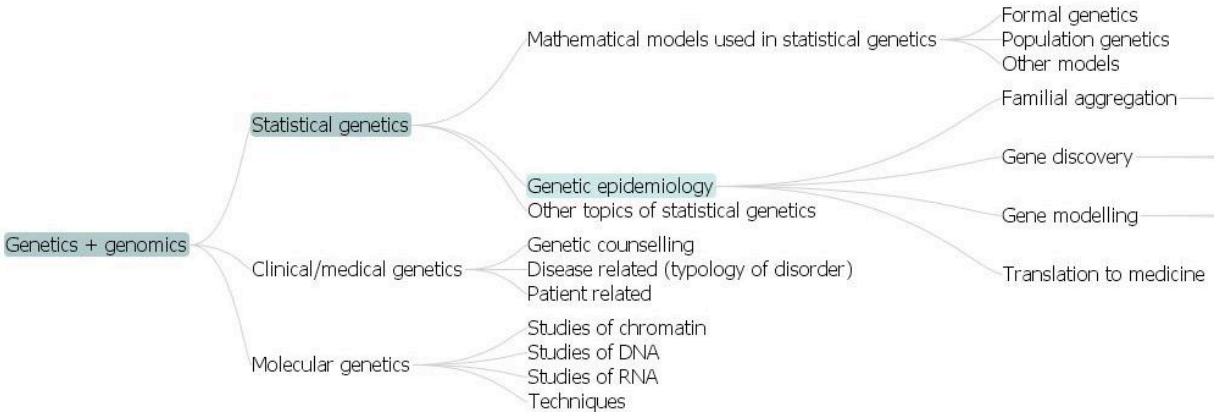

Figure 3. Part of the theme hierarchy

Therefore, each resource has associated metadata composed of: a list of ontology concepts and their frequencies found automatically for textual resource or manually for images and videos; a set of free text keywords, title and author's name; text from the abstract; a set of themes/topics it addresses; the original language of the resource and reference to available translations (if applicable); and the academic level for which the content is suitable.

A part of the content submission page is shown in Figure 4.

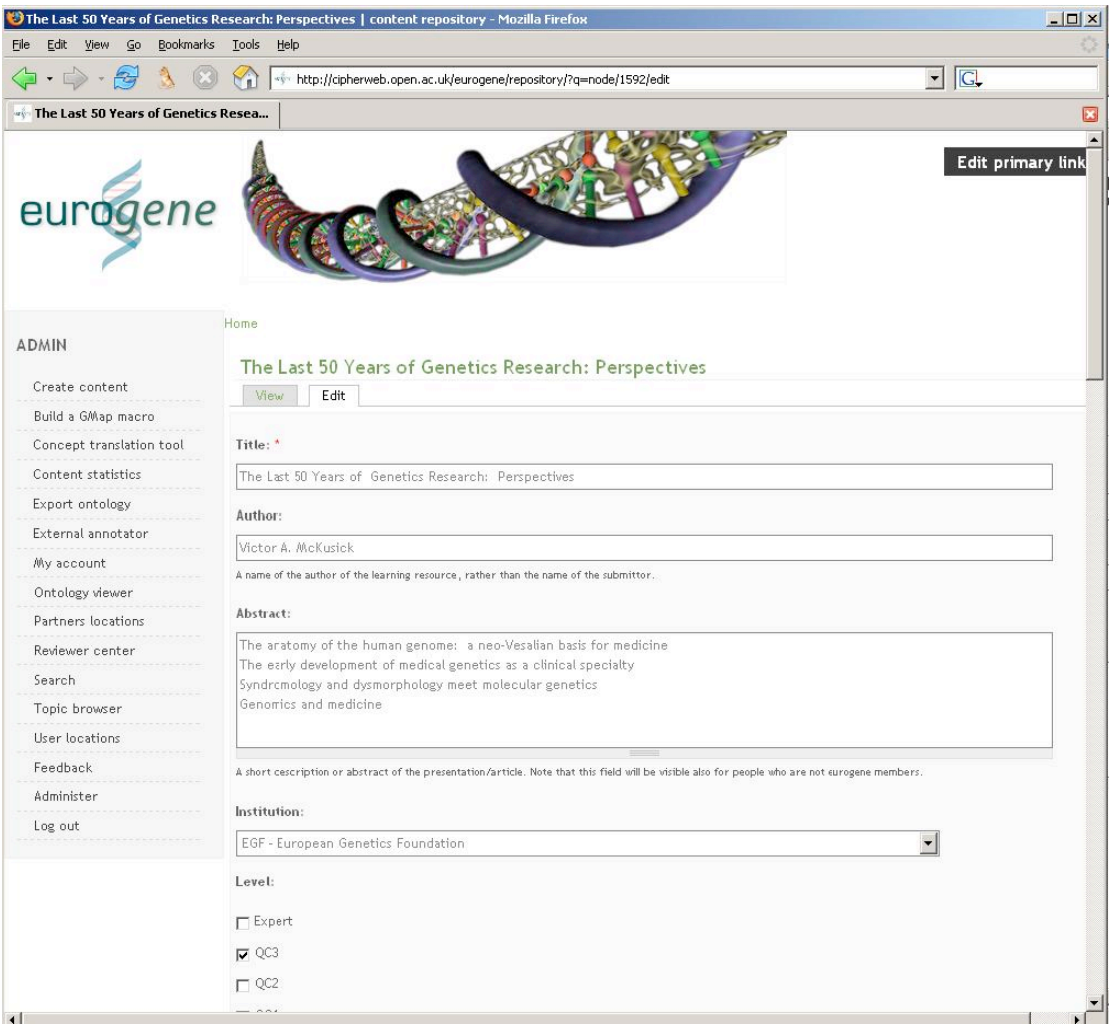

Figure 4. Eurogene submission page

The resources are also represented in a SKOS (Simple Knowledge Organization System) like format, i.e. each resource is associated with a list of ontology concepts independently of the original language.

4 Similarity measures

The annotation of resources makes it possible for search algorithms to combine the concept matching based on semantic similarity with traditional string matching. We define three different measures for similarity based concept matching:

- Identity – verbal expressions (term) of two different educational resources are considered equal if the corresponding concepts are identical, i.e. if they belong to the same synset defined across all supported languages.
- Generalisation – for concept matching, two concepts count as equal, if one is a generalisation of the other and their distance in the hierarchy is shorter that a predefined threshold parameter, and
- Common hypernym – for concept matching, two concepts count as equal, if they share a common hypernym and their edge distance in the tree is shorter than a predefined parameter.

These three cases of concept matching are shown in Figure 5. Term 1 and Term 2 can be in different languages. If the distance parameter is greater than 3, then concepts C1, C3 in b), but not C1 , C4 in c) count as equal. If its value is 4 then the concepts both in b) and c) are considered as equal.

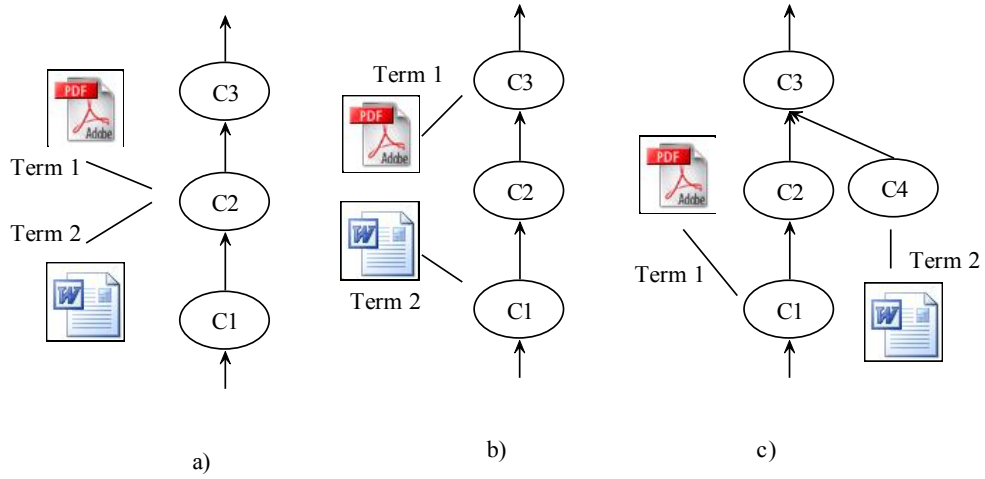

Figure 5. Terms describing the same concept (a) and two different concepts (b, c).

Applying measures according to b) and c) requires additional heuristics to resolve the situations where the same concept might be included more that once. In the rest of the paper we will use only the first measure of concept matching.

5 Search and reasoning

Annotated educational resources can be explored in different ways. Simple search allows the user to request documents from the repository. The query may be multilingual, expressed as a Boolean function of the following metadata types: terms, topics, text field, author's name, title, words in the Abstract or in the text. Any Boolean function constructed by AND, OR, NOT operators and parentheses is allowed. For example, the query "linkage[term]" AND "marcador genético[term]" AND ("Genetic epidemiology[topic]" OR ("Dawn Teare[author]" AND "estimate[text]")) combines two terms, in two different languages, a topic/theme, author's name and free text. Terms, topic/theme and the author's name are auto-completed from the ontology, theme hierarchy and the Eurogene repository. The results can be further filtered by language, academic level and resource type. An example of the filter specification is shown in Figure 6.

Figure 6. Query filter

Concepts found in the educational resources and their frequencies are used to calculate the similarity between these resources either as the correlation or as cosine similarity. For example, a lecture from the Göttingen University course of genetic epidemiology has been annotated with the result shown in Table 2. The left column are concepts from the ontology, the right column is their absolute frequency.

Ontology concept	frequency
segregation analysis	21
gene	9
dominant	7
inheritance	6
recessive	4
population	4
locus	3
allele frequency	3
allele	3
…	…

Table 2. Annotation of a lecture

Similarity between two annotated resources is calculated as Pearson's correlation coefficient r_{xy} as follows:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}.$$

In this formula x_i and y_i are concept frequencies in compared resources.

The similarity measure allows us to specify queries by providing an example of a resource (scenario 1 - QBE). In practice, it means that the learner, who does not fully understand the topic described in some resource can use this resource as a query to the system. The query resource and the answer could be in different languages. Similarity measures can also be used to find pathways from the initial resource, which is specified as knowledge prerequisites, to the final learning resource characterised by the target knowledge of the course (scenario 2). Finally, learners may be interested to find a resource that discusses the same topic in less or more depth. This scenario has been tested in a different domain (scenario 3). These scenarios will be demonstrated in the case study presented in the next section.

6 Case study

The case studies we present make use of the content from the Göttingen University and Université Paris-Sud. The one semester course in statistical genetics from Göttingen, consisting from 12 lectures with PowerPoint presentations in English provided us with the test data. A part of the annotation table is shown in Table 3.

	Lecture 1	Lecture 2	Lecture 3	Lecture 4	Lecture 5	Lecture 6	Lecture 7	Lecture 8	Lecture 9	Lecture 10	Lecture 11	Lecture 12
genetics	20	23					1				3	
chromosome	7	2			4		3		1	2		1
affected	5	1	7			17	1	11	3		4	
population	5		2	4	1			4	2		15	3
cell	4		2									
allele	3	16	2	3	9	11	4	6	10	9	7	10
gene	3	8	2	9	4		4		1	4	2	
genotype	3	3	2	2	5	1	1	1	6	2	7	13
inheritance	3	1	2	4	1	1		3	3	1	1	
…												

Table 3. Annotation of Göttingen lectures

These presentations were classified into the themes as shown in Figure 7.

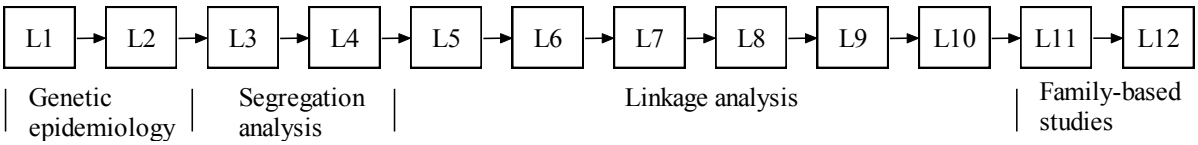

Figure 7. Themes of the Göttingen course

In this case study we show Scenario 1 (Query By Example) and 2 (Educational Paths) as described in Section 1. Scenario 3 cannot be demonstrated because of small number of resources participating in this case study.

6.1 Query by Example

To demonstrate Query By Example (QBE), we compare three annotated educational resources with the set of Göttingen presentations to find out if there is a similar one. The comparison will be based on the Pearson's correlation coefficient and we say that two resources are semantically similar if $r_{xy} \geq 0.5$. For testing we used PowerPoint presentations from Université Paris-Sud. Each onewas submitted as a "query" and the task was to find out whether there is a semantically similar presentation within the Göttingen course.

As the first "query" we used the presentation entitled "Family-Based Tests in Inbred Populations". The maximum value of Pearson's correlation coefficient was $r_{xy} = 0.5956$ for Göttingen presentation L11. All remaining Göttingen presentations gave significantly lower values (around 0.1 to 0.2). This result confirms the naïve expectation, based on the comparison of the presentation title and theme associations shown in Figure 7.

The second example used as a query was the presentation entitled "The HapMap project". This large international project investigated genetic similarities and differences in human beings. From the title of the presentation it was not possible to judge what part of genetics is addressed. The maximum value of Pearson's correlation coefficient was $r_{xy} = 0.2$, most values were even lower that 0.1. The conclusion is that none of the Göttingen presentations addresses similar content.

The third example query was the presentation in French, entitled "Analyse de liaison génétique pour les maladies multifactorielles". The Pearson's correlation coefficient gave two close high values $r_{xy} = 0.58$ and $r_{xy} = 0.59$ for Göttingen presentations L2 and L1, respectively.

All these results were presented to domain experts who found them satisfactory and approved them.

6.2 Educational Paths

The task is to organise the educational content into a sequence which is a pathway from the initial state describing the learner's knowledge prerequisites to the final state characterising the goal of the learning process. The pathway can be understood as the way of organising educational material into a learning package for a course. The initial state could be specified by concepts that the learner should know prior to taking the course. The final state could be also specified by a set of concepts the learner should understand after taking the course. However, in this case study we specify the initial state by the presentation selected for the lecture. Semantic cohesion between presentations will again be measured in terms of the Pearson correlation coefficient. The matrix of correlation coefficients is shown in Table 4.

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12
L1	1	0.475	0.094	0.111	0.101	0.142	0.194	0.138	0.188	0.070	0.116	0.082
L2		1	0.074	0.143	0.257	0.083	0.134	0.173	0.203	0.237	0.351	0.165
L3			1	0.775	0.044	0.072	-0.031	0.055	0.008	0.083	0.030	0.013
L4				1	0.063	0.030	0.004	0.032	0.026	0.144	0.044	0.025
L5					1	0.220	0.433	0.303	0.390	0.096	0.180	0.139
L6						1	0.161	0.914	0.557	0.234	0.155	0.116
L7							1	0.220	0.285	0.041	0.067	0.056
L8								1	0.641	0.230	0.193	0.181
L9									1	0.164	0.135	0.266
L10										1	0.096	0.045
L11											1	0.216

Table 4. Correlation matrix

In the matrix, rows and columns denote learning resources, the elements of the matrix are correlation coefficients between the corresponding resources. The correlation matrix is symmetrical which means that the correlation coefficient do not define the direction in the "knowledge space". Due to the symmetry, correlation coefficients alone cannot be used to decide whether a sequence leads from the simple to the difficult or vice versa. However, this is frequently the case of similarity measures, mutual information/entropy based criteria are also symmetrical. It only means that the direction must be defined outside of this conceptual framework.

The task can be now formulated as finding the sequence of learning resources that optimises a criterion calculated from the correlation matrix. There are multiple strategies for organising the pathway. In this case study we decided to maximise the sum of correlation coefficients along the learning path. It means that we want to organise learning resources in a way that maximises average correlation between adjacent presentations. Therefore, we are looking for an algorithm that construct a resources Li, starting from the selected initial resource, each resource is in the sequence exactly once and the sum of corresponding correlation coefficients is maximum. Let's assume that the initial lecture is the same as in the original course, i.e. L1. Solution for the Göttingen course presentations is shown in Figure 8.

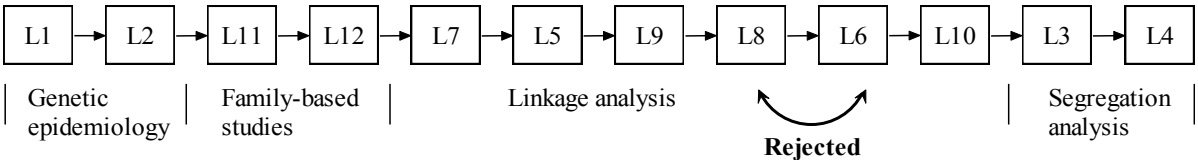

Figure 8. Reordered Göttingen course

The sum of correlation coefficients is 4.568. When comparing the sequence in Figure 8 with the original one in Figure 7, we can see that the new sequence reorganises presentations within one theme and reorganises the sequence of themes, but the presentations remain within the original themes. The most striking difference is that two themes, Family-based studies and Segregation analysis, have been swapped. Moreover, within the Linkage analysis theme the sequence of presentations is also changed. These proposed changes have been discussed with the author of the Göttingen course and she accepted all changes with only one exception: presentation L8 must (!) follow L6. We have analysed this requirement and calculated the matrix for corrected sequence. The sum of correlation coefficients is 4.480, i.e. 2% lower than the maximum value. For comparison, the sum for the original sequence is 3.102.
The case study has been presented only to demonstrate the approach. The presented results certainly do not allow us to draw any serious conclusion. However, the described framework proposes the methods that help the content users to exploit fully the content repository, allow them to evaluate their choice and search for alternatives.

7 Related work and conclusions

The semantic annotation of educational resources can be used not only for querying the resource repository by metadata, but also for reasoning across multiple resources [4], [5]. Automatic annotation of learning resources and assembling learning objects has been used for instance in [3]. Educational pathways have been studied for some time by observing how learners exploit resources on the web. For example, [1] define so called Walden's Paths as a tool that "employs metadocuments to superimpose structure over unconnected documents to facilitate their reuse via coherent presentations". Resnik in [10] proposed information-based model for assessing semantic similarity, Levene and Loizou in [6] developed a probabilistic model based on Markov chains for evaluating sequences of web-base resources. In [8], pathways are used for post-visit exploration of museum documents.

The task of finding an optimal sequence is isomorphic with the travelling salesman problem, which is known to be computationally complex (NP-complete). Finding a solution by brute force does not scale up. The size of the problem presented in the above case study was close to the limits of available computers. At present (i.e. in September 2009), the Eurogene repository contains 1,152 learning resources and is growing every day. Using the same approach for the whole repository would not be possible. However, the Eurogene knowledge base offers additional structures, such as a theme hierarchy and associations between themes and ontology concepts, that can be used for constructing heuristics to factorise and prune the search space.

Eurogene offers not only tools that help the content user to construct educational pathways, but also supports the development of learning packages and their full integration with the content of the repository. Machine translation runs in the background when new content is submitted. The SYSTRAN machine translation software needs to be trained for a specific domain, in the case of Eurogene it is genetics. After being trained, the quality of machine translation is acceptable. However, the software does not resolve various problems with the layout of some presentations, especially if the length of sentences in the source and target languages differs significantly.
Maintenance of the Eurogene ontology an important issue for the sustainability of the repository. The ontology must be updated because as new content is submitted it may reveal existing gaps and the domain itself is rapidly expanding. The current approach combines the calculation of potential gaps based on the evaluation of the available content with the assessment by a panel of experts.

References:

[1] Furuta R., Shipman F.M. and Wilson H. (2002). Metadocuments as Communicative Artifacts to Enable Use of a Research Digital Library in Undergraduate SMET Education. *http://www.csdl.tamu.edu/walden/reports/wp-project-description.pdf,* accessed June 2009

[2] ISO 2788-1986. Documentation - Guidelines for the establishment and development of monolingual thesauri. *International Organisation for Standardization (ISO*). British equivalent BS 5723: 1987, new version BS 8723-1: 2005

[3] Jovanovic J., Gasevic D. and Devedzic V. (2006). Ontology-Based Automatic Annotation of Learning Content. *Int. Journal on Semantic Web & Information Systems*. 2(2). Pp. 91-119. April-June 2006.

[4] Knolmayer G.F.(2003). Decision support models for composing and navigating through e-learning objects. In HICSS '03*: Proceedings of the 36th Annual Hawaii International Conference on System Sciences*. page 31.3, Washington, DC, USA, 2003. IEEE Computer Society.

[5] Kontopoulos,E., Vrakas,D., Kokkoras,F., Bassiliades, N. and Vlahavas I.(2008). An ontology-based planning system for e-course generation. *Expert Syst. Appl.,* 35(1-2):398–406

[6] Levene M. and Loizou G. (2003). Computing the Entropy of User Navigation in the Web. International Journal of Information Technology and Decision Making 2(3): 459-476

[7] Miller G.A. (1998). Nouns in WordNet. In *WordNet An Electrionic Lexical Database (C. Fellbaum, ed.),* The MIT Press. Cambridge Mass. 23-67

[8] Mulholland, P., Collins, T. & Zdrahal, Z. (2005). Bletchley Park Text: using mobile and semantic web technologies to support the post-visit use of online museum resources. *Journal of Interactive Media in Education* (Portable Learning: Experiences with Mobile Devices. Special Issue).

[9] Ontologies (2009):
Gene Regulation, http://www.gene-regulation.com/pub/databases.html, accessed March 2009,
Mammalian Phenotype, http://www.informatics.jax.org/searches/MP_form.shtml, accessed March 2009,
Gene Ontology, www.geneontology.org/*,* accessed March 2009,
NCI Thesaurus, http://nciterms.nci.nih.gov/*,* accessed March 2009
Unified Medical Language System *(*UMLS*),* http://www.nlm.nih.gov/research/umls/, accessed March 2009

[10] Resnik P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130

[11] WordNet (1998).*An Electronic Lexical Database* (C. Fellbaum ed.). MIT Press

Authors:

Zdenek Zdrahal
Petr Knoth
Trevor Collins
Paul Mulholland
The Open University, Knowledge Media Institute
Walton Hall, Milton Keynes, MK7 6AA, UK
{Z.Zdrahal; P.Knoth; T.D.Collins; P.Mulholland}@open.ac.uk